# Annotating a large corpus with anaphoric links

Agnès Tutin, François Trouilleux, Catherine Clouzot, Éric Gaussier, Annie Zaenen, Stéphanie Rayot, Georges Antoniadis

# Annotating a large corpus with anaphoric links

**Agnès Tutin[+], François Trouilleux[°], Catherine Clouzot[+], Eric Gaussier[°], Annie Zaenen[°], Stéphanie Rayot[°+], Georges Antoniadis**

[°]Xerox Research Centre Europe
6, chemin de Maupertuis. 38240 Meylan, France
{trouilleux;gaussier;rayot;zaenen}@xrce.xerox.com

[+]Equipe CRISTAL-GRESEC,
Université Stendhal - Grenoble 3, BP 25
F-38040 Grenoble Cédex 9
{tutin;clouzot;antoniadis}@u-grenoble3.fr

**Abstract**

This paper presents a one million word French corpus annotated with anaphoric links. The anaphoric expressions selected are mainly grammatical discourse phenomena for which a reliable annotation could be provided. The annotation scheme, defined in XML, encodes the orientation of the anaphoric relation by using a specific element for relating the anaphoric expression to its antecedent(s). A set of five semantic relations is used to type the anaphoric relation. As a rule, linguistic expressions selected are phrases, but the annotation scheme uses specific elements to deal with descriptive anaphors which occur in nominal ellipses and demonstrative anaphors. Special cases such as multiple antecedents, discontinuous elements or ambiguity are discussed.

## 1. Introduction

Building corpora with anaphoric links is essential in NLP and in linguistics. In NLP, it may enable one to design robust anaphora resolution techniques and realistic strategies to generate referring expressions. Nevertheless, to our knowledge, few corpora with anaphoric links are available to date, probably because the task is time-consuming and because achieving a high inter-annotator-agreement seems difficult.

In this paper, we present a project of a one million word French corpus annotated with anaphoric links. We first outline the methodological choices we made in selecting the anaphoric expressions, delimiting the linguistic expressions and coding the anaphoric relations. We then present in detail the annotation scheme. We finally briefly compare our project to other annotation schemes.

## 2. Methodological choices

While building our annotated corpus with anaphoric information, a set of methodological choices had to be made: What kind of anaphoric elements should be selected? How should one delimit the linguistic elements involved in the anaphoric relation? How should the anaphoric relation itself be encoded? What technical choices to make (markup language, preprocessing and editing tools)? In the study described here, these choices were made under strong time pressure - since the coding schema had to be developed and the one million word corpus had to be annotated and checked in 12 months. Moreover creating an annotated corpus implies that the annotation can be done reliably and quickly: annotated corpora are meant for broad use and cannot cost too much, so the distinctions made should be easy to understand and quick to make.

### 2.1. Anaphoric expressions selected

Dealing with every kind of anaphoric and cataphoric expression in a one million word corpus is an unfeasible task. In this project, two main criteria were used to select a valid subset of anaphoric expressions:

- The wish to deal mainly with **discourse phenomena**, more than with purely syntactic intrasentential phenomena. We thus discarded the study of reflexive pronouns – always coreferential with the subject in French – and relative pronouns governed by well known syntactic rules.
- The **feasibility of the task**: proposing a consistent annotation throughout a large corpus was a major preoccupation and we chose to exclude anaphoric phenomena for which this guarantee could not be given. This was the case for non-elliptical definite NP's. Dealing with this kind of anaphoric expressions appeared premature given the lack of satisfactory formal description and hence the likelihood that the annotators' decisions would be highly subjective, as has been highlighted by Poesio and Vieira (1998). Some anaphoric pronouns, adjectives and adverbs (*cela, ça, là, ici, tel, etc.* "that, there, here, such, etc.") pointing back to "indistinct" antecedents (Corblin, 1987) were also ruled out, as well as verbal ellipses for which an informal evaluation showed that they were almost always missed by annotators.

The anaphoric and cataphoric expressions finally retained were elements involving closed classes, whatever the syntactic nature of the antecedent (NP,AP,VP, clauses, sentences). More specifically:

- Third person anaphoric personal pronouns, to the exclusion of reflexive pronouns,

```
1. Le législateur ne s'est pas risqué à
```

définir ce concept. Sans doute n'a-t-**il** pas voulu, en agissant de la sorte, figer cette notion ...

The legislator has not taken the risk to define this concept. In doing so, most likely he has not wanted to fix the notion...

- Possessive pronouns and determiners,

2. Et, faute d'un véritable travail de recherche sur la recherche, chacun y va de **son** exemple, de **ses** a priori, ou de **ses** intérêts, pour défendre **son** point de vue.

   And, in the absence of a real research activity about research, everybody uses his example, his a priori, or his interests to defend his point of view.

- Demonstrative anaphoric pronouns, except "neuter" pronouns (*ce, ça, cela, ceci*) ( " this, that ")

3. Mais la croissance de l'emploi n'a pas empêché **celle** du chômage.

   But the growth of employment has not prevented that of unemployment.

- Indefinite pronouns, including compounds with nominal heads, such as *un ensemble, la plupart*, ("a set, most") and numerals,

4. Parmi les projets qui s'adressent à l'ensemble de la RFTS, **certains** selon leur nature pourront éventuellement continuer à être gérés comme projets communs.

   Among the projects that concern the whole of the RFTS, certain [ones] according to their nature could potentially continue to be managed as common projects.

- The "proverb" *le* + *faire*, (" do [it] ")

5. Nous souhaiterions toutefois, dans l'intérêt de la défense des intérêts moraux des chercheurs, que la jurisprudence accueille plus favorablement qu'elle ne **le fait** actuellement leur action [...]

   We would wish, however, that, in the interest of the defense of the moral interests of the researchers, the jurisprudence would be more favorable than it is now to their action [...]

- Anaphoric adverbs such as *dedans, dessus*, ("in, on")

6. Quelqu'un d'autre a cru à son idée, a travaillé **dessus** et a réussi à faire aboutir le vaccin.

   Somebody else believed in his idea, worked on [it] and succeeded in making the vaccine.

- Nominal ellipses.

7. [...] l'originalité d'une œuvre se teinte différemment d'un type d'œuvre à **l'autre ∅**.

   [...] The originality of a work has a different aspect from one type of work to another.

- Anaphoric "pointers" belonging to a closed class: *ce dernier, le premier*, etc. ("this last [one], the first [one]") when they cannot be analyzed as including nominal ellipses.

8. Même si l'apport créatif des auteurs scientifiques porte plutôt sur le contenu, les idées, il n'en demeure pas moins que **ces derniers** disposent d'une certaine liberté pour les exprimer de manière originale.

   Even when the creative contribution of the scientific authors is more on the side of the content, the ideas, it remains true that the latter have a certain freedom to express them in an original way.

## 2.2. Principles used for delimiting the antecedent and anaphoric/cataphoric expressions

The first task to deal with when building a corpus with anaphoric annotations is to identify the linguistic elements involved in this discourse relation consistently. The anaphoric/cataphoric element and its antecedent(s)[1], must be isolated in the running text.

The main problem we encountered in the delimitation task is the fact that many linguistic elements involved in anaphoric relations do not correspond to traditional phrases, in particular in non-coreferential anaphoric relations (see 3.2.4.). For example, in noun head ellipses, the anaphoric element is the ellipsis itself, not the NP it is part of , while the antecedent is generally a noun, not a full NP. In the following extract, the noun-head ellipses, here marked with "∅", point back, the one to *électrification* and, the other one to *sous-stations* (the anaphoric relation is indicated with a subscript letter).

9. On a pour (la deuxième **électrification**$_a$) (21 **sous-stations**$_b$), là où pour (la première **∅**$_a$) il en aurait fallu (35 **∅**$_b$).

   One finds for the second electrification 21 substations where 35 were needed for the first.

[1] We will use this term for both anaphoric and cataphoric relations.

Both ellipses are part of NPs and point back to nouns that are part of NPs. If one wanted to strictly annotate the anaphoric expressions, one would not mark NPS, but only the ellipses and nouns. Because we wanted to ensure compatibility with syntactic annotations and we did not want to lose information about the discourse relation[2], we mark the expressions on two levels:

- the anaphoric element and its antecedent(s) are annotated at the level of the constituent they are internal to. In our example, all bracketed elements will be retained.
- the segments in the antecedent (if any) which are not taken by the delimited anaphoric noun phrase will be specifically marked up as such.

In most cases, namely coreference anaphoric relations, the syntactic level and the anaphoric level coincide.

## 2.3. Coding the anaphoric relation

The anaphoric markup should indicate: a) the elements involved in the anaphoric relation, i.e. which anaphoric expression is related to which antecedent(s), b) the discourse/semantic relationship between the anaphoric expression and its antecedent(s).

### 2.3.1. Relating an anaphoric expression to its antecedent(s)

An anaphoric expression cannot be autonomously interpreted, but needs an antecedent to get a referential/semantic content. As a consequence, the relations we want to encode may be seen as oriented from the anaphoric expression to its antecedent. This differs from coreference relations that are generally seen as symmetrical.

This orientation was a major difference with other annotation schemes mainly designed to encode coreferential relations, for example in the MUC (Chinchor and Hirschmann 1997) or in the MATE (Davies *et al.* 1998) projects, a difference which led us to constrain the linking mechanism in our annotation scheme.

### 2.3.2. Types of anaphoric relations

Since our project involves a large set of anaphoric expressions and extends the scope of antecedents to all kinds of phrases (not only NPs, but also APs, clauses and sentences), we encountered other types of anaphoric relations than coreference. These discourse relations had to be encoded with a small set of semantic tags, easy to use and likely to meet a high inter-annotator agreement. This led us to define five classes, which could be further refined in several subclasses.

**Coreference**

If the anaphoric expression denotes the same discourse referent as its antecedent, the type of the relation is "coreference".

10. BP vend sa branche détergents ménagers et produits de toilette.

    BP is selling its branch of household detergents and toilet articles.

**Set membership**

If the anaphoric expression denotes a referent which is an element or a subset of the referent denoted by its antecedent, the anaphoric relation is of type "set membership". We do not differentiate between an element-set relation on the one hand and a subset-superset relation on the other. The "set-membership" relation implies that the antecedent expression denotes a set.

In the following example, *l'une* ("the one") and *les trois autres* ("the three other ones") denote respectively an element and a subset of the set denoted by *Des quatre locomotives de Savoie* ("the four locomotives of Savoie"). In both cases, the relation is of type "set-membership".

11. **Des quatre locomotives de Savoie, l'une** est à redresseurs [...]. **Les trois autres** montrent une sorte de coexistence.

    Of the four locomotives of Savoie, one is of the erector type [...]. The three others show a kind of coexistence ...

The "set-membership" relation is also used in cases where the antecedent denotes a class and the anaphoric expression an instance of that class, e.g.

12. **Le lion** est peut-être un grand chasseur, mais **celui que Pierre a tué** n'était pas dangereux.

    The lion might be a mighty hunter but the one that Pierre killed was not dangerous.

**Description**

Various theories of reference differ in the way they associate objects in the universe of discourse with linguistic descriptions. For example, in *Marie est intelligente et Jeanne l'est aussi* ("Marie is intelligent and Jeanne is it too"), we consider that the VP, *est intelligente*, has no referent and thus the anaphoric relation between this VP and *le* is of the type "description". This distinction between referential and non-referential expressions led us to establish anaphoric relations of type "description" in either of two situations. If neither the antecedent nor the anaphoric expression are referential expressions (*i.e.* neither of them denotes a referent, they only describe one), the anaphoric relation is of type "description". In the following sentence, the antecedent of the clitic pronoun *l'* is *exploitées* ("exploited"), an expression which does not denote anything, but only describe the referent denoted by *l'énergie hydraulique* .

[2] We tried to make sure that the annotation scheme would encode all the information necessary to the study of discourse mechanisms in descriptive anaphoric relations involved in noun head ellipses and demonstrative anaphors.

13. [...] si toutes les ressources énergétiques naturelles sont **exploitées**, l'énergie hydraulique **l'** est insuffisamment.

    [...] while all natural energy sources are exploited, hydraulic energy is [it] insufficiently.

If both the antecedent $exp_i$ and the anaphoric expression $exp_j$ are referential expressions and are neither linked by a coreference or set-membership relation and if the description that $exp_i$ provides of its referent is needed in $exp_j$ to identify its referent, the anaphoric relation is of type "description". In the following example, the expression *la première génération* ("the first generation") describes its referent as being of type génération. This description also applies to the referent of *la deuxième* ("the second") and is needed to identify this referent.

14. [...] **la première génération** est celle des locomotives des débuts jusqu'aux années 1930 , **la deuxième** étant celles des machines transformées [...]

    [...] the first generation is that of locomotives from the beginning to the thirties, the second being that of transformed engines [...]

The delimitation of the two expressions in this example is justified in a following section (section 3.2.3.). Note that the "set membership" relation also implies a "description" relation. Nominal head ellipses and demonstrative pronouns will sometimes be linked to there antecedent by a "set-membership" relation, sometimes by a "description" relation.

**Sentential antecedent**

When the antecedent of an anaphoric expression is a clause or a sentence, we consider that the anaphoric relation is of type "phrase", even if it could have been annotated either as type "coreference" or type "description". In particular, when antecedents can be analyzed as indirect speech clauses, like in the following example, one can consider that only the textual content is pointed back to:

15. **Ces records se déroulent**, il faut **le** dire, **dans une période exceptionnellement favorable à l'innovation technique ferroviaire en France**.

    These records take place, one has to admit, during a period that is exceptionally favorable for technical innovations conerning railroads in France

However, in practice, it proved difficult with some verbs to differentiate clause antecedents with an indirect speech status from usual clause antecedents. To avoid inconsistency in the tagging process, we finally opted for the more neutral relation "phrase".

**Indefinite relation**

Finally, we distinguish a fifth type of anaphoric relation meant to cover all cases not covered by the four previous types. An example of such a relation is when anaphoric expression is negatively quantified:

16. Parmi ces étudiants, **aucun** n'a fait son travail..
    Among these students, none has done his work.

## 2.4. Technical choices

As a markup language, we chose to use the standard XML, even if this language, though powerful, can appear cumbersome insofar as it necessitates specific editing tools and parsers. For the tagging process itself, we used lighter proprietary formats that were transformed in standard XML.

The texts we had to markup were provided by ELRA in a TEI Lite format. Whenever possible, we tried to adopt the TEI guidelines for our specific markup (see 3.1.).

The annotation process was performed by hand by two skilled linguists (a Master's student and a Ph.D. student). A large subset of anaphoric expressions was automatically pre-annotated. Antecedents and anaphoric relations had to be marked up manually, but editing tools were used to make the task easier. The participants in the project met regularly to discuss the problems encountered by the annotators, e.g. problems delimiting antecedent boundaries, determining the type of the anaphoric relation or dealing with ambiguities.

## 2.5. Evaluation

In order to evaluate the quality of the annotation, we have decided that each annotator should review the annotation made by the other. Each annotator will be supported in this task by different members of the project, so as to have an as objective as possible evaluation. This evaluation will be performed on at least 5% of the corpus, i.e. approximately 50,000 words. Detailed measures, including precision and recall, will be provided, based on the following typology of errors: missing anaphoric expressions (anaphoric expressions not annotated), spurious anaphoric expressions (expressions that have been wrongly considered as anaphoric), interpretation errors (anaphoric expressions linked to a wrong antecedent), antecedent delimitation errors (the antecedent is correctly identified, but some elements should be added or removed), anaphoric expression delimitation error (as before but for anaphoric expressions), link type errors (the link used between an anaphoric expression and its antecedent is not correct), and others.

Even though we do not expect all these error types to be encountered in equal proportions in the corpus, our typology should be sufficiently fine-grained to help us identify where remaining problems are (if any), and make appropriate decisions.

Lastly, in addition to standard evaluation measures, we will compute the inter-annotator reliability on an

independent subset of the corpus, consisting of approximately 20,000 words. This last measure should help us understand further the degree of the homogeneity of the annotation and will serve as a complement to standard measures for future uses of the corpus.

## 3. The annotation scheme

The annotation scheme is defined in XML[3] . The texts are divided into sections, paragraphs (<p>) and sentences (<s>). Sections and paragraphs were marked up in the original corpus. The segmentation of the texts into sentences was done using the XRCE natural language processing tools. Only the elements we introduced to describe anaphoric relations will be presented here. They all appear below the sentence level.

### 3.1. Overview

This section introduces the main aspects of the annotation scheme. The presentation distinguishes three aspects:

- delimiting expressions,
- linking expressions,
- typing the relation between expressions.

Expressions and anaphoric links are marked up using separate elements, in a style close to what is recommended in the MATE project[4].

#### 3.1.1. Delimiting expressions

**General case.**

Expressions which are either anaphoric or the antecedent of an anaphoric expression are annotated as <exp> elements. Every <exp> element has an attribute named "id", the value of which is of type ID (i.e. a unique identifier in the document).

17. **<exp id="f17">BP</exp>** vend **<exp id="f18">sa</exp>** branche détergents ménagers et produits de toilette.

    BP is selling its branch of household detergents and toilet articles.

**Discontinuous antecedents.**

The antecedent of an anaphoric expression can be discontinuous, in particular when a comment clause is inserted in another clause[5]. In such cases, the two segments of the antecedent expression are annotated as two separate <exp> elements which are linked together with the attributes next and prev provided by the TEI[6]. The values of next and prev are of type IDREF. The first <exp> element has the attribute next="X" where X is the value of the id attribute of the second element. The second <exp> element has the attribute prev="Y" where Y is the value of the id attribute of the first element[7].

In the following example, the antecedent of the pronoun *le* is *Un programme de rachat a été élaboré afin de diminuer la charge financière de la dette publique*, a clause which is discontinous due to the presence of the inserted comment clause *on le sait* ("as is known", lit: "one knows it").

18. **<exp id="f28" next="f30">Un programme de rachat a été élaboré</exp>**, on <exp id="f29">le</exp> sait, **<exp id="f30" prev="f28">afin de diminuer la charge financière de la dette publique</exp>**

    A buy-back program has been elaborated, [one knows it], to decrease the financial charges of the public debt.

#### 3.1.2. Linking expressions

In our scheme, describing anaphoric links consists in linking <exp> elements together. Among the solutions proposed by the TEI to link expressions together, we have choosen to use <ptr> ("pointer") elements[8]. A <ptr> element specifies a relation from a place in the document (the place where the <ptr> element appears in) to one or several elements of the document, by means of an attribute (called "target" in the TEI) the value of which is of type IDREFS. A <ptr> element is an empty element.

The link between an anaphoric expression and its antecedent(s) is thus indicated with a <ptr> element placed immediately after the opening <exp> tag delimiting the anaphoric expression. The antecedent(s) of the anaphoric expression are identified with an attribute called "src" (corresponding to the TEI "target" attribute), whose value is the value of the id attribute of the antecedent <exp> element.

19. **<exp id="f17">BP</exp>** vend **<exp id="f18"><ptr src="f17"/>sa</exp>** branche détergents ménagers et produits de toilette.

    BP is selling its branch of household detergents and toilet articles.

The src attribute may have several values, separated by a white space, when the anaphoric expression has several antecedents[9].

**20.** Bonn estime que **<exp id="f3">le président de la Commission européenne</exp>** "n'a pas tenu compte" des suggestions que lui avait faites **<exp id="f5">le chancelier Kohl</exp>** lors de **<exp id="f6"><ptr type="coref" src="f3 f4"/>leur</exp>** dernière

[3] http://www.w3.org/XML

[4] section 4.1 “Markup declaration”

[5] This situation is distinct from cases where an anaphoric expression has multiple antecedents; see below.

[6] TEI, section 14.7 "Aggregation".

[7] The two attributes are redundant, but they allow one to reconstruct the antecedent from any of its part.

[8] TEI, section 6.6 "Simple Links and Cross References".

[9] When anaphoric expressions point back to several sentences, this is considered as a standard case of multiple antecedents (each sentence is marked up separately and the src attribute has several values).

rencontre.

Bonn thinks that the president of the European Commission did not take into account the suggestions that Chancellor Kohl had made during their previous meeting.

In some rare cases, delimiting an antecedent appears difficult, even when the anaphoric status is unquestionable. This is often the case with "adverbial" personal pronouns *y* and *en*, which point back to a part of the text that cannot be easily identified. These pronouns often have a summarizing function:

21. Quoi qu'il **en** soit...

Be this as it may...

As it might be interesting to distinguish these cases from the non-referential uses of adverbial pronouns and to calculate statistics about their respective distribution, we decided not to skip these cases. That is why, though being unable to locate an antecedent, we indicate these pronouns with the help of a specific empty element <ptr-i> (for "indefinite pointer"), inserted in the <exp> element. For example, for our previous example, we would have:

22. Quoi qu'il **<exp id=="f4"><ptr-i/> en</exp>** soit.

Be this as it may

#### 3.1.3. Type of the relation

The type of the relation between an anaphoric expression and its antecedent(s) is indicated with an attribute called "type" in the <ptr> element. The value of a type attribute may be one of "coref" (coreference), "mde" (set-membership, "membre de" in French), "desc" (description), "phrase" (sentential antecedents) or "indef" (indefinite relation).

23. **<exp id="f50">Des quatre locomotives de Savoie</exp>**, **<exp id="f51"><ptr type="mde" src="f50"/>l'une</exp>** est à redresseurs [...].**><exp id="f52"><ptr type="mde" src="f50"/>Les trois autres</exp>** montrent une sorte de coexistence ...

Of the four locomotives of Savoie, one is of the erector type [...]. The three others show a kind of coexistence ...

### 3.2. Conventions for the delimitation of antecedents and anaphoric expressions

We present here the conventions that the delimitation of the expressions linked in an anaphoric relation obeys. As a rule, as outlined in 2.2, standard phrase boundaries are marked. When some elements within the phrases are not specifically involved in the anaphoric relation, they are annotated as such.

We first present the general convention for delimiting antecedents and anaphoric expressions, then the convention for linking expressions in a relation of type "description" or "set membership". This latter convention will lead us to introduce a new element (<seg>).

#### 3.2.1. Delimitation of antecedents

**Identification.**

As a rule, we retain as the antecedent of an anaphoric expression an expression which is:

- non-pronominal,
- and as close as possible to the anaphoric expression.

The first constraint will lead to the following style of annotation, where the three anaphoric expressions point to the same antecedent:

24. <s> Si **<exp id="f35">la CGT</exp>** pousse à l'élargissement, <exp id="f36**"><ptr type="coref" src="f35"/>elle**</exp> ménage en même temps l'opinion publique. </s> <s> C'est ainsi qu'<exp id="f39**"><ptr type="coref" src="f35"/>elle**</exp> a marqué <exp id="f40"><**ptr type="coref" src="f35"/>ses</exp>** réserves face au blocage de voies [...]. </s>

Although the CGT pushes towards extension, it tries to treat public opinion carefully. For this reason it has expressed its reservations with respect to the road blocks [...].

25. Dès **<exp id=e1><ptr type="coref" src="e2"/>sa</exp>** naissance, **<exp id="e2">le réseau Internet</exp>** a échappé aux réseaux qui **<exp id="e3"><ptr type="coref" src="e2"/> l'</exp>**avaient financé.

From its birth on the Internet has escaped [the control of those] that financed it.

The reason for putting the constraint that the antecedent be a non-pronominal expression is twofold: one the one hand, we want to make clear that the relation we encode is oriented from one anaphoric expression to a more specific one; on the other hand, this convention allows one to remain as independent as possible from presuppositions about the structural information at the sentence level which might influence the interpretation of pronouns. In our view, linking pronominal expressions to non-pronominal ones allows one to focus on the interpretation, regardless of the procedure of interpretation.

In some cases, the antecedent can be a non-anaphoric pronoun, in particular when the pronoun has a generic human interpretation, like in the following example:

26. Ces expériences permettent à **<exp id="e1">chacun </exp>**de remettre en cause ou d'affiner **<exp id="e2"><ptr type="coref" src="e1"/> sa</exp>** vision

du monde (ou de l'entreprise).

These experiments allow everybody to question or to refine his vision of the world or the enterprise.

More rarely, the anaphoric expression cannot be linked to a non-pronominal phrase, because it specifically refers back to another anaphoric expression. In this case, the value for the SRC attribute is the ID value of the anaphoric antecedent involved.

27. Les relations entre **<exp id="e1">pays plus ou moins développés</exp>** prendraient alors un tour nouveau,**<exp id="e2"><ptr type="mde" src="e1"/> les uns</exp>** vendant des idées qu' **<exp id="e3"><ptr type="coref" src="e2"/> ils</exp>** n'ont pas encore eues ...

    The relations between more or less developed countries would take a new turn, the ones selling ideas that they have not yet had…

**Antecedent boundaries**

As a rule, antecedents are phrases.

NPs include restrictive modifiers, i.e. noun adjuncts that play a role in the identification of the referent. Attributive APs, restrictive relative clauses, restrictive PPs will thus be included in the antecedent NP, as in the following example:

28. <s>Faire rouler un train en traction diesel en 1939 est certes faire de l'État un encaisseur de taxes, mais c'est aussi tabler sur **<exp id="e1">un mode de traction qui est loin d'avoir atteint <exp id="e2"><ptr type="coref" src="e1"/> sa</exp> maturité technique</exp>**. </s>

    Getting a train with a diesel engine to run in 1939 is certainly to allow the State to collect taxes but it also means to rely on an engine and drivers section that is far from having reached [its] maturity.

As a side effect, an anaphoric expression can be included in the antecedent NP, as in the last example.
On the other hand, if modifiers are used to add information about the referring expression, but are not used to delimit the reference, they will not be part of the NPs. In the following example, the appositive relative clause has not been included in the source NP.

29. **<exp id="e5">L'usine marnaise</exp>**, qui appartient au groupe Beghin-Say, produit annuellement environ 80 000 tonnes de sucre blanc. **<exp id="e6"><ptr type="coref" src="e5"/>Elle</exp>** ...

    The Marne factory, which belongs to the Beghin-Say group, produces annually around 80,000 tons of white sugar. It…

When appositions are proper nouns, they are included in the antecedent, since it appeared hard to decide for one or the other NP as an antecedent[10].

30. **<exp id="f15">Le PDG de Peugeot, M. Jacques Calvet</exp>**, s'est vanté d'avoir roulé en BX à plus de 200 km à l'heure, et **<exp id="f16"><ptr type="coref" src="f15"/>il</exp>** plaide pour la liberté de la vitesse sur autoroutes.

    The CEO of Peugeot, Mr. Jacques Calvet, has boasted that he had driven in a BX at more than 200 km per hour and pleads for speeding freedom on the superhighways.

As sentential antecedents and infinitive clauses are sometimes hard to delimit, it was decided to select the largest possible antecedent.

Other antecedent phrases - APs, PPs - did not present any specific problems and are thus delimited in the conventional way.

### 3.2.2. Delimitation of anaphoric expressions

Anaphoric expressions are generally easy to delimit. As a rule, phrases including the anaphoric element are annotated. In most cases, the phrase only includes the anaphoric element.

Modifiers determining anaphoric pronouns are included in the anaphoric elements.

31. **<exp id="e14"><ptr type="coref" src="e13"/>Eux aussi </exp>** se sont avérés capables de paralyser les usines pour exprimer **<exp id="e15"><ptr type="coref" src="e13"/> leurs</exp>** mécontentements.

    They too have succeeded in paralysing factories to express their dissatisfaction.

32. Dans <exp id="e12">le département qui nous occupe plus particulièrement </exp>, <exp id="e13"><ptr type="desc" src="e12"/>**celui de la fabrication du yaourt**</exp>, aucun ouvrier de production n'a dépassé le niveau de l'école primaire.

    In the department that concerns us more particularly, that of the production of yogurt, no production worker has gotten above the grade school level.

The modifiers that are not involved in the anaphoric relation will be isolated with a <seg> element, as will be shown below.

[10] In these cases, we did not achieve a satisfactory inter-annotator-agreement and decided thus to include both expressions.

NPs containing noun ellipses will be annotated as anaphoric NPs, though it can be argued that only the ellipsis is the anaphoric marker.

33. L'informatisation de la production [...], dans <exp id="e1">un cas</exp>, prend le nom de commande numérique et, dans **<exp id="e2"><ptr type="desc" src="e1"/>l'autre</exp>**, de pilote automatique.

    The computerization of the production [...] is in one case called ‘çommande numérique’ and in the other ‘pilote automatique’.

For possessive determiners, only the anaphoric expression will be annotated, though the determiner can be considered as a kind of modifier. In the following example, *son* in *son autonomie* ("her autonomy") can be analyzed as ***l**'autonomie **d'elle***. ("the autonomy of her")

34. <s><exp id="e1">La traction</exp> perdait ainsi <exp id=e2><ptr type="coref" src="e1"/>**son**</exp> autonomie. </s>

    In this way, the engine lost its autonomy.

For "proverbs" *le* + *faire*, the pronoun *le*, which cannot be dissociated from *faire*, was included in the anaphoric expression, as are negative adverbs when they occur.

35. <s> Si nous <exp id="e11"><ptr type="desc" src="e10"/>**ne l' avons pas fait** </exp>plus tôt, c'est que notre démarche construisait un raisonnement ... </s>

    If we didn’t do this earlier, it is because our thinking has constructed a way of reasoning...

### 3.2.3. Delimitation of expressions in relations of type “description”

If there is an anaphoric relation of type "description" between two referring noun phrases, we have chosen to annotate the complete noun phrases rather than just the antecedent description and the pronoun. Given *la croissance de l’emploi… celle du chômage* (" the growth of employment ... that of unemployment " ), we annotate:

36. Mais **<exp id="f41">la croissance de l'emploi</exp>** n'a pas empêché **<exp id="f42"><ptr type="desc" src="f41"/>celle du chômage</exp>.**

rather than:

37. Mais la **<exp id="f41">croissance</exp>** de l'emploi n'a pas empêché **<exp id="f42"><ptr type="desc" src="f41"/>celle</exp>** du chômage.

    But the growth of employment has not prevented that of unemployment.

This practice is justified by several reasons. We make use of a "set-membership" relation which involves a relation between referents and- the delimitation of referring expressions; the same type of expressions will be found as anaphoric in a relation of type "description"; so it is consistent to annotate them in the same way throughout the corpus.

One may also note that if one of the expression linked in a relation of the type "description" were to be the antecedent of a coreferring pronoun, the annotation scheme would require only the delimitation of the referring expression, as marking both the referring expression and the description would lead to a structure that is unnecessarily complex.

### 3.2.4. Distinctive descriptions

However, we wanted to distinguish precisely, in the two expressions linked by a relation of type "description" which parts described the two referents and which part described only one of them. For this reason, we introduced a `<seg>` element, with an attribute `type` valued `"distinctif"`, which is used to delimit the segment in an antecedent $exp_i$ which only applies to the referent of $exp_i$ and not to the referent of the anaphoric expression $exp_j$ it is the antecedent of. In anaphoric relations of this type, there is usually in the antecedent and in the anaphoric expression some modifier(s) which distinguishes the two referents. The `<seg>` tags aims at delimiting the segment in the antecedent NP, which is specific to the antecedent. Such segments will usually be adjectival phrases, prepositional phrases or relative clauses[11]. A complete annotation of the example above then would be:

38. Mais **<exp id="f41">la croissance <seg type="distinctif">de l'emploi</seg></exp>** n'a pas empêché **<exp id="f42"><ptr type="desc" src="f41"/>celle du chômage</exp>**

    But the growth of the employment has not prevented that of unemployment

## 3.3. Special cases.

This section introduces the annotation conventions for a few special cases: double anaphoric links and sloppy identity, ambiguities, coordinations, bound anaphors.

### 3.3.1. Double anaphoric link

The possessive pronouns (*le sien, la sienne, le leur*, etc.;" his, hers, theirs") involve a double anaphoric link: a link of type "description" and a link of type "coreference". In the sentence *Pierre préfère la fille de Jeanne à la*

[11] The idea of delimiting a distinctive description in the antecedent is similar to the notion of “repudiation” proposed by Halliday and Hasan, (1976): "In any anaphoric context, something is carried over from a previous instance. What is carried over may be the whole of what there was, or it may be only part of it; and if it is only part of it, then the remainder, that which is not carried over, has to be REPUDIATED." (p. 93)

*sienne* (" Pierre prefers the daughter of Jeanne to his own "), the possessive pronoun *la sienne* denotes "Pierre's daugther". This interpretation requires (1) that the description *fille* ("daugther") be inferred from *la fille de Jeanne* ("the daughter of Jeanne") and (2) that the referent identified as a *fille* be identified as linked by a possessive relation to the referent of *Pierre*.

Cases of "sloppy identity" are analogous. In the sentence *L'homme qui donne son salaire à sa femme est plus sage que celui qui le donne à sa maîtresse*, (" the man who gives his salary to his wife is wiser than the one who gives it to his mistress " ) interpreting the clitic pronoun *le* involves inferring the description *salaire* (" salary " ) and a possessive relation between the referent of *le* and the referent of *celui qui le donne à sa maîtresse (*" the one who gives it to his mistress'' *).*

Such situations always involve an anaphoric link of the type description and they are the only cases where a link of this type can be viewed as involving two antecedents. Taking advantage of this observation, we will annotate such phenomena as anaphoric links of type `desc` with two values for the `src` attribute: the first value identifies the expression where the needed description is to be found, the second value identifies the expression which denotes the possessor.

```
39. Marie aime <exp id="f1">la fille de
    Jeanne</exp>; <exp id="f2">Pierre</exp>
    préfère <exp id="f3"><ptr type="desc"
    src="f1 f2"/>la sienne</exp>.

    Marie likes the daughter of Jeanne;
    Pierre prefers his own.

40. L'homme qui donne <exp id="f1">son
    salaire</exp> à sa femme est plus sage
    que <exp id="f2">celui qui <exp
    id="f3"><ptr type="desc" src="f1
    f2"/>le</exp> donne à sa
    maîtresse</exp>.

    The man who gives his paycheck to his
    wife is wiser that the one that gives it
    to his mistress.
```

### 3.3.2. Ambiguities

When the interpretation of an anaphoric expression is ambiguous (*i.e.* the annotator cannot chose between several possible antecedents), the annotation scheme allows the use of multiple `<ptr>` elements. In the following sentence, the clitic pronoun *le* may either be interpreted as denoting "Mary's salary" (in which case the anaphoric link is of type "coreference" or "Jeanne's salary" (in which case it is an instance of sloppy indentity). The ambiguity preserving annotation will be:

```
41. Marie dépose <exp id="f1">son
    salaire</exp> à la banque et <exp
    id="f2">Jeanne</exp> <exp id="f3"><ptr
    type="coref" src="f1"/><ptr type="desc"
    src="f1 f2"/>le</exp> dépense aussitôt.

    Marie puts her salary in the bank and
    Jeanne spends it immediately.
```

It should be noted that the annotator should use multiple `<ptr>` elements only when he or she cannot identify an antecedent for certain, not for cases when he or she might think that there might be some structural ambiguity as would appear with automatic anaphora resolution systems, for instance.

A second type of ambiguity occurs when the annotator is unable to decide wether the expression is anaphoric or not. We encountered several cases of this type with demonstrative pronouns which can have either a generic or an anaphoric interpretation. In the following example, *ceux* can either be understood as a generic human referent ("all the people") or as a subset of "the specialists". In this case, we introduce a specific attribute `st` (for "status") which takes the value `"incertain"` (uncertain).

```
42. <s> <exp id="e1">Les spécialistes</exp>
    remarquaient cependant que le franc
    restait ferme face au dollar et à la
    livre. </s><s> Quant à <exp id
    ="e2"><ptr type="mde" src="e1"
    st="incertain"/>ceux qui craignaient de
    voir la chute du billet vert pénaliser
    les valeurs d'exportation</exp>, <exp
    id="e3"><ptr type=coref"
    src="e2"/>ils</exp> ouvraient là un
    débat qui n'a pas encore été tranché ...

    The specialists noticed, however, that
    the frank remained stable with respect
    to the dollar and the pound. As far
    those that feared that the fall of the
    greenback would penalize the export
    values, they opened a debate that has
    not yet been decided ...
```

### 3.3.3. Anaphoric *en*

There are in French some cases of anaphora with the clitic *en* where the anaphoric expression may be interpreted as composed of two disjoint segments, e.g. *en* and *deux* in the sentence *Pierre a trois enfants; Marie en connait deux*("Pierre has three children; Marie has two [of them/that type of entity]"). In our annotation scheme, the two expressions are annotated separately, the relation between the two expressions being viewed as a syntactic phenomena which is out of the scope of the scheme.

```
43. Augmenter <exp id="f20">un emprunt</exp>
    coûte normalement moins cher à un
    débiteur que d'<exp id="f21"><ptr
    type="desc" src="f20"/>en</exp> lancer
    <exp id="f22"><ptr type="desc"
    src="f20"/>un nouveau</exp>.

    To increase a loan costs a debtor
    normally less to get a new one.
```

### 3.3.4. Conjoined antecedent NPs

Antecedent NPs can be coordinations. In this case, should we delimit each NP within the conjoined NP or only consider the whole NP? We opted for the first solution, arguing that NPs included in coordinations could be antecedents of an anaphoric expression. Conjoined antecedent NPs are simply analyzed as a specific case of

multiple antecedents (see 3.1.2.). However, the conjoined NP was not given a status of <exp>[12], but annotated with a <seg> element with an attribute type="coord".

```
44. [...] <seg type="coord"><exp id="e1">les
    embryologistes</exp> et<exp id="e2">les
    neurobiologistes de deux laboratoires de
    l'Institut Pasteur associés au
    CNRS</exp></seg> ont créé une souche de
    souris mutante, insensible à la nicotine
    . Pour ce faire,  <exp><ptr type="coref"
    src="e1 e2"/>ils</EXP> ont inactivé
    [...]

    [...] the embryologists and the
    neurobiologists of the two laboratories
    of the Institut Pasteur that are
    associated with the CNRS have created a
    kind of mutant mice that is not
    sensitive to nicotine. To do this they
    have inactivated [...]
```

#### 3.3.5. Bound anaphors

In our annotation scheme, bound anaphors do not receive any special markup. In most cases, they can be analyzed as involving coreference relations, as in the following example:

```
45.   <exp id="e3932">chacun</exp> est
    libre d'effectuer des copies des
    oeuvres dont <exp id="e3934"><ptr
    type="coref" src="e3932"/>il</exp> a
    besoin.

    Everyone is free to make copies of
    the works he needs.
```

## 4. Comparison with other annotation schemes

In this section, we will briefly compare our annotation scheme to three other existing annotation systems :

- The UCREL Discourse Annotation scheme (Garside *et al.* 1997),
- The MUC Annotation scheme (Chinchor & Hirschmann 1997),
- The MATE Annotation scheme (Davies *et al.* 1998).

### 4.1. Linguistic expressions selected

Our annotation scheme exclusively deals with anaphoric phenomena, in particular grammatical anaphoric expressions. The Lancaster/IBM (UCREL) project was more ambitious since it aimed at annotating all kinds of anaphoric relations, including bridging anaphors. The first objective of the MUC annotation scheme was to build a reference corpus for the MUC information task. It only deals with coreferential relations, in the broad sense, including all types of NPs. The MATE project proposes two schemes: a) a core scheme only dealing with coreferential NPs and b) an extended scheme including all kinds of anaphors, including bridging anaphors (though the authors seem skeptical about the feasibility of the annotation task for this kind of phenomena).

### 4.2. Types of relations between linguistic elements.

Our annotation scheme proposes five types of anaphoric/cataphoric relations that have been tested on our corpus.

The MUC scheme only deals with coreference relations, as in the core MATE scheme. The extended MATE scheme allows more relations (bound anaphors, function-value, element-set, subset-set, attribute-of, part-of, "strict possession", instantiation, event relation, situation) which, to our knowledge, have not been tested on a large scale. The anaphoric/cataphoric relations of the UCREL scheme are in accordance with Halliday and Hasan's analysis (1976): REF (coreference), SUBST (substitution), ELL (ellipses), IMP (implied anaphora), OF (NP with inferrables *of*-complement), predicative, MISC (miscellanous), META (metatextual reference). Most relations, widely illustrated in Halliday and Hasan's study, seem easy to use, though IMP and OF relations seem to overlap in some cases.

As in the UCREL annotation scheme, we opted for a restricted set of anaphoric relations, but we did not consider the ellipses as a kind of relations, but as (empty) anaphoric expressions. We were led to introduce an *ad hoc* type "phrase", to avoid inconsistency with sentential antecedents.

### 4.3. Linking the linguistic elements

Most systems use internal links to relate the linguistic elements, i.e. annotations on linguistic elements that point to other linguistic elements. In our annotation scheme, we use a specific empty element (<ptr>), inserted in the anaphoric/cataphoric expressions.

In the MUC scheme, links between coreferential elements are noted by means of a SGML attribute REF pointing to the ID of a coreferring expression (since coreference is symmetric and transitive, an expression can point to any other coreferring expression).

```
46. <s> <COREF ID="0">Ocean Drilling &
    Exploration Co.</COREF> will sell <COREF
    ID="3" MIN="business"><COREF ID="2"
    TYPE="IDENT" REF="0">its</COREF>
    contract-drilling business</COREF>, and
    took a $50.9 million loss from
    discontinued operations in <COREF
    ID="12" MIN="quarter">the third
    quarter</COREF> because of the planned
    sale. </s>
```

As we already mentioned, the linking of expressions is further constrained in our scheme to mark the orientation of the relation from the anaphoric expression to its antecedent.

In the MATE meta-scheme (Poesio, 1999), the linking of expressions is done in a stand-off annotation style, with a <link> element pointing to one of the expressions in

[12] Contrary to the MATE annotation scheme, where every NP is annotated, e.g. :
<de ID="40"><de ID="41">John</de> and <de ID="42">Louise</de></de> went to ...

relation, and containing an <anchor> empty element pointing to the other expression. As far as the linking of expressions is concerned, our <ptr> element may be seen as equivalent to the MATE <anchor> element, with this restriction that the <ptr> element, rather than being placed in a <link> element pointing to an element $e_i$, is placed immediately after the opening tag of the element $e_i$ itself. The annotation in 47 in our scheme is equivalent to the MATE style annotation in 48:

47. `<exp id="X">PCDATA</exp> <exp id="X"><ptr src="X"/>PCDATA</exp>`

is equivalent to the following MATE style annotation:[13]

48. `<de id="X">PCDATA</de> <de id="Y">PCDATA</de> <link href="Y"><anchor href="X"/></link>`

Stand-off markup in XML documents tends to be unreadable for a human annotator without the help of some user interface. The inclusion of the <ptr> element at the level of the anaphoric expression greatly facilitate the annotator's work. We think, even though we can at present not provide a formal correspondence, that our linking scheme contains all the information required for a translation into a format compliant with the MATE guidelines.

The UCREL scheme takes advantage of the asymmetry of the anaphoric relation. In this light proprietary format, rich in linguistic information, the anaphoric relation is coded on the anaphoric expression by means of an identifier related to the antecedent. In the following example, the two *he* are linked to the antecedent *Gagnon* by "2", the "<" indicates an anaphoric relation and "REF" is used to type a coreferential relation.

49. `(2 Gagnon 2) said later <REF=2 he approved of the penalties... and that <REF=2 he considers the case closed.`

This annotation scheme provides a complete markup system for special cases such as multiple, ambiguous or uncertain antecedents, but the proprietary format could not probably be easily transformed in a standard markup language such as XML.

## 5. Conclusion

Annotating anaphoric relations in a large corpus proved a feasible task insofar as 1) we excluded complex anaphoric expressions (such as "neuter" demonstrative pronouns) and b) we chose a simple annotation scheme involving few anaphoric relations. Non coreferential anaphoric expressions involving sentential antecedents or nominal ellipses could be annotated even if the expressions boundaries appeared sometimes hard to delimit.

The annotation task was mostly performed by hand given the lack of any available training corpus including anaphoric information. The results provided by our corpus could enable to partially automate the tagging process in the future, though discarding non referential pronouns or locating ellipses seem far from being straightforward tasks.

[13] Taking <exp> as equivalent to <de> and allowing some incompleteness in the use of the two annotation schemes.

## References


Chinchor N., Hirschmann L. (1997), MUC-7 Coreference Task definition, Version 3.0, *Proceedings of MUC-7*. http://www.muc.saic.com

Corblin F. (1987), *Indéfini, défini et démonstratif*. Genève, Droz.

Davies S., Poesio M., Bruneseaux F., Romary L., (1998), *Annotating Coreference in Dialogues : Proposal for a Scheme for MATE (First Draft).*

Garside R., Fligestone S. & Botley S. (1997), Discourse annotation : anaphoric relations in corpora, in R. Garside, G. Leech & A. McEnery (eds), *Corpus annotation : Linguistic Information from Text Corpora*, London, Longman.

Halliday M. & Hasan R. (1976) *Cohesion in English*, London, Longman.

Poesio M. & Vieira R., (1998), A corpus-based investigation of definite description use. *Computational Linguistics*, **24**, 2.

Poesio, Massimo (1999) MATE Dialogue Annotation Guidelines – Coreference. Second draft. http://www.ims.uni-stuttgart.de/projekte/mate/mdag/cr/cr_1.html


## Annex A : DTD used in the annotation scheme

The DTD presented here does not account for father nodes of the <p> elements, but makes use of a fake <text> element as the root of the XML tree.

```
<!ELEMENT text (p+)>
<!ELEMENT p (s*)>
<!ELEMENT s (#PCDATA|exp|seg)+>
<!ELEMENT exp (ptr*,ptr-i*,(#PCDATA|seg|exp)+)>
<!ELEMENT ptr EMPTY>
<!ELEMENT ptr-i EMPTY>
<!ELEMENT seg (#PCDATA|exp)+>

<!ATTLIST p     type CDATA #IMPLIED >
<!ATTLIST p     n CDATA #IMPLIED >
<!ATTLIST p     nom CDATA #IMPLIED >
<!ATTLIST p     id ID #IMPLIED >

<!ATTLIST exp   id ID #REQUIRED
                next IDREF #IMPLIED
                prev IDREF #IMPLIED>

<!ATTLIST ptr   type
(coref|mde|desc|phrase|indef) #REQUIRED
                src IDREFS #REQUIRED
                st (incertain) #IMPLIED>

<!ATTLIST seg   type (distinctif|coord)
#REQUIRED>
```

## Annex B : An example of annotated text

<s> **<exp id="e131">L'expression oeuvre scientifique</exp>**, objet de notre étude ne se laisse pas facilement appréhender par le droit. </s><s> On peut **<exp id="e132"><ptr type="coref" src="e131"/>lui</exp>** donner un sens très général et considérer que l'expression vise toute production intellectuelle de caractère scientifique (§ 1. ). </s><s> Il est possible de **<exp id="e133"><ptr type="coref" src="e131"/>lui</exp>** donner un contenu plus restreint si l'on met l'accent sur le terme oeuvre (§ 2. ).</s>

<s> **<exp id="e134">Le mot oeuvre</exp>** a, en droit, plusieurs significations. </s><s> Selon le vocabulaire juridique de l'association Henri Capitant dirigé par le Doyen Cornu, **<exp id="e135"><ptr type="coref" src="e134"/>il</exp>** revêt notamment les sens suivants : </s>

<s> ouvrage résultant d'une construction (immobilière) ; </s>

<s> activités déployées en vue d'un but déterminé (activités de l'entreprise ou activités universitaires et sociales). </s>

<s> D'une manière générale, **<exp id="e136"><ptr type="coref" src="e134"/>il</exp>** s'analyse comme le résultat d'un travail ou d'une activité manuelle ou intellectuelle. </s><s> À l'évidence, c'est cette dernière acception qui semble la plus adaptée pour notre étude. </s><s> Précisément en quoi consiste les résultats du travail du scientifique (I) ? </s>
<s> Après avoir répondu à cette question, on s'attachera à cerner les caractéristiques de l'activité scientifique (II). </s>

<div3 n="1.1.1.1" id="BO1-1.1.1.1">

<s> Les résultats de l'activité scientifique </s>

<s> Tenter de définir aujourd'hui la science en général et l'activité scientifique en particulier semble une démarche vouée à l'échec tant est vaste le champ de l'activité scientifique. </s><s> Le constat est fait par **<exp id="e137">les scientifiques <exp id="e138"><ptr type="coref" src="e137"/>eux-mêmes</exp>**</exp> ainsi que par de nombreux philosophes des sciences. </s><s> La science d'aujourd'hui écrit M. Kourganoff " est une réalité complexe dont il est difficile de donner une définition générale ". </s><s> Il paraît, en revanche, plus facile d'indiquer en quoi consistent les résultats de la recherche scientifique. </s>

<s> Les résultats de l'activité scientifique vont dépendre du type de recherche en cause : **<exp id="e139">recherche fondamentale</exp>** ou **<exp id="e140">recherche appliquée</exp>** ; **<exp id="e141"><ptr type="coref" src="e139"/>l'une</exp>** est tournée vers la science pure, l'explication du réel, **<exp id="e142"><ptr type="coref" src="e140"/>l'autre</exp>** vers la technique, c'est-à-dire l'action sur le réel. </s><s> " Les fins, les voies, les démarches des deux recherches fondamentale ou appliquée, ne sont pas semblables. </s>
<s> " La remarque contient à l'évidence une certaine proportion de vérité. </s>
<s> On a toutefois exagéré à l'excès la distinction entre ces deux types de recherche.
</s>